APLICACIÓN DE MÉTODOS DE CLASIFICACIÓN AL DOWNSCALING ESTADÍSTICO

Rafael Cano (1)
Francisco J. López (2)
Antonio S. Cofiño (3)
José Manuel Gutiérrez (3)
Miguel A. Rodríguez (4)

(1) SED del CMT en Cantabria y Asturias. INM
(2) Servicio de Modelización Numérica del Tiempo. INM
(3) Dept. Matemática Aplicada, Universidad de Cantabria
(4) Instituto de Física de Cantabria (CSIC / Univ. de Cantabria)

RESUMEN

En este artículo se analizan diversas alternativas para la predicción probabilística de meteoros utilizando el método de análogos, y se presentan resultados de validación en la red principal de estaciones del INM en las diversas cuencas peninsulares. En primer lugar se comparan distintas especificaciones (tamaño de la rejilla y rango horario) del vector 4D que define el "estado de la atmósfera" en base a las predicciones de un modelo numérico. Seguidamente, se describe la forma más conveniente de comprimir esta información, eliminando redundancias, utilizando componentes principales. A continuación, se analizan distintas técnicas de clasificación para obtener un conjunto de estados de la atmósfera análogos a uno dado, de entre aquellos disponibles en una base de datos de estados históricos (reanálisis). Finalmente, se muestra la forma de mejorar la resolución de una predicción realizada por el modelo numérico a partir de la estadística de los meteoros locales observados en las fechas correspondientes a los análogos hallados para el estado definido por la predicción.

1. Introducción

Un problema de gran interés práctico en el ámbito de la predicción meteorológica es la adaptación a escalas regionales o locales de las salidas sobre una rejilla de los modelos numéricos de circulación atmosférica (como el HIRLAM o el modelo operativo del CEPPM); este problema se denomina "*mejora de resolución*" (downscaling) y es de gran importancia en diversos sectores productivos que necesitan predicciones de gran resolución en áreas locales (predicciones urbanas, riesgo de heladas, predicción estacional de cultivos, etc). La forma más simple de adaptar las salidas a puntos concretos consiste en interpolar los puntos de rejilla más cercanos, pero este sistema no aporta ningún detalle nuevo a la predicción. Por ello, hasta la fecha se han aplicado diversas técnicas más sofisticadas para resolver este problema, que pueden clasificarse en (ver Zorita y Storch (1997) para una detallada descripción): *Técnicas dinámicas*, que utilizan las salidas del modelo numérico como condiciones de contorno de un modelo de mayor resolución y parametrizaciones físicas apropiadas (p.ej., el modelo HIRLAM utilizado en el INM), y *técnicas estadísticas*, que combinan las predicciones en rejilla del modelo numérico con la información estadística de mayor resolución contenida en los registros históricos que estén disponibles en el área de interés (por ejemplo, los tomados diariamente en la red de observatorios del INM). En la figura 1 se ilustra la diferencia entre la resolución típica de un modelo (la rejilla del CEPPM a 1ºx1º) y la dada por la red secundaria de observatorios del INM.

La aplicación de estas últimas técnicas ha sido impulsada recientemente por la disponibilidad de bases de datos de predicciones desarrolladas en diversos proyectos de reanálisis (p.ej., reanálisis ERA-15 del ECMWF, que incluye las salidas diarias de un mismo modelo para el período 1979-1993); estas bases de datos sirven de puente entre las salidas de los modelos numéricos y los registros históricos locales de meteoros disponibles, permitiendo aplicar técnicas de regresión, correlación, etc. para combinar la información. Entre las técnicas que han mostrado ser más eficientes desde el punto de vista operativo destaca el método de análogos, que opera en dos etapas (ver Lorenz, 1969; Toth, 1990):

- Primero, dada una predicción realizada por el modelo numérico, se obtiene un conjunto de configuraciones de la atmósfera análogas, de entre aquellas disponibles en una base de datos (reanálisis).

- A continuación, se realiza la predicción local de un meteoro concreto a partir de la estadística dada por sus sucesos en las fechas correspondientes a los análogos hallados.

Un estudio comparativo de distintos métodos de downscaling, incluyendo una versión simple del método de análogos, ha sido recientemente publicada por Zorita y Storch (1999). La eficiencia de los distintos métodos basados en análogos depende en gran medida del algoritmo específico utilizado para la selección de análogos. Hasta la fecha, el algoritmo más utilizado ha sido el de los *k*-vecinos, consistente en seleccionar los *k* días más similares al día problema utilizando una cierta métrica. En este trabajo analizamos diversas técnicas de agrupamiento que pueden ser aplicadas a este problema (*k*-medias, mapas auto-organizativos, etc.) y mostramos su eficiencia frente al método de *k*-vecinos considerando el Brier Skill Score obtenido al predecir la precipitación, la racha de viento máxima y la insolación en la red de estaciones completas del INM. Para ello se utilizarán las predicciones del modelo operativo CEPPM como patrones representativos de la configuración atmosférica del día problema. A continuación describimos en mayor detalle los datos utilizados.

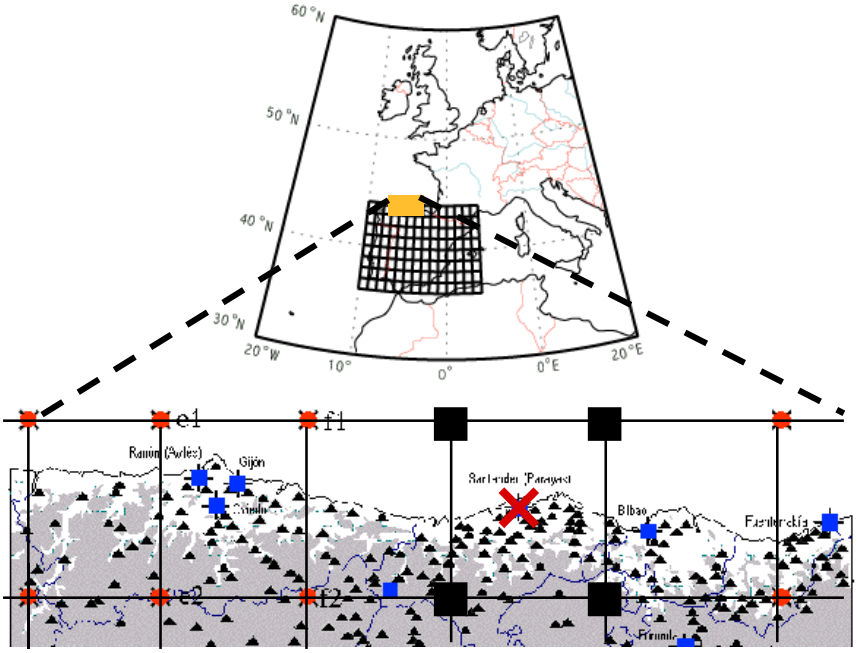


Figura 1: Ilustración del problema de la mejora de resolución de la predicción. La resolución de la rejilla de salida (la península Ibérica a 1ºx1º en la parte superior) no es suficiente para precisar detalles regionales. La figura inferior muestra un área magnificada (la cornisa Cantábrica) con el detalle de todas las estaciones de la red secundaria del INM (triángulos), junto con la rejilla del modelo numérico.

2. Descripción de los Datos y Eliminación de Redundancia

Uno de los requisitos principales para realizar un downscaling estadístico es mantener la consistencia del modelo numérico operativo (del que se obtendrán las predicciones) y del modelo integrado en el reanálisis para crear la base de datos. En esta sección describimos los datos y modelos utilizados en este trabajo.

2.1 Reanálisis ERA y Salidas del Modelo Operativo

Como base de datos para realizar la búsqueda de análogos hemos utilizado el reanálisis ERA-15 del CEPPM (`http://www.ecmwf.int/research/era/index.html`), que integra un modelo T106L31 proporcionando valores diarios de Temperatura (T), Humedad relativa (H), Geopotencial (Z) y componentes U, V del viento en seis niveles (300, 500, 700, 850, 925, y 1000 mb) a las 00, 06, 12 y 18 UTC, el período comprendido entre los años 1979 y 1993 (el reanálisis proporciona otras variable y niveles, pero se ha tratado de mantener la dimensión limitada para evitar un coste computacional excesivo). El modelo de reanálisis es una versión simplificada del modelo operativo T511L60 del CEPPM; este modelo nos proporciona los pronósticos de las variables anteriores necesarios para llevar a cabo el método de análogos. Al utilizar dos modelos similares en cuanto a su concepción y

parametrizaciones mantenemos la consistencia necesaria entre la base de datos y las predicciones operativas.
Un problema que no ha sido suficientemente abordado en la literatura es la especificación de un patrón óptimo para la configuración atmosférica para un problema dado (elección de la región de la rejilla, y las variables, niveles y horas apropiados); obsérvese que la información proporcionada por el modelo numérico resulta excesiva y redundante para una manipulación eficiente; por otra parte, dependiendo del problema puede resultar óptimo considerar una región concreta y unas variables, niveles y horas apropiadas. En este trabajo el área geográfica de interés se limita a la península Ibérica. Por tanto, hemos restringido los campos del reanálisis y los operativos para trabajar en esta zona, considerando dos alternativas posibles para la especificación del estado de la atmósfera:

- *Modelo 1 (**M1**)*: Rejilla limitada al área peninsular de 1ºx1º de resolución y estática en el tiempo para una hora de predicción dada (12 UTC) (ver figura 2(a)). En este caso, el estado de la atmósfera está caracterizado por el vector de estado:

$$\mathbf{U} = (T_{12}^{1000}, ..., T_{12}^{300}, H_{12}^{1000}, ..., H_{12}^{300}, ..., V_{12}^{1000}, ..., V_{12}^{300}) \qquad [1]$$

  Donde X_i^j denota el nivel de presión *j* de la variable X a las *i* UTC. Por tanto, en este caso cada vector tiene 135(nodos)x5(variables)x6(niveles) = 4050 dimensiones.

- *Modelo 2 (**M2**)*: Una rejilla limitada geográficamente a cada cuenca peninsular e incluyendo las distintas salidas horarias disponibles para un mismo dia (00, 06, 12, 18 y 24 UTC) (ver figura 2(b)):

$$\mathbf{V} = (\mathbf{U}_{00}, \mathbf{U}_{06}, \mathbf{U}_{12}, \mathbf{U}_{18}, \mathbf{U}_{24}) \qquad [2]$$

  En este caso cada vector de estado tiene 25x5x6x5 = 3750 dimensiones.

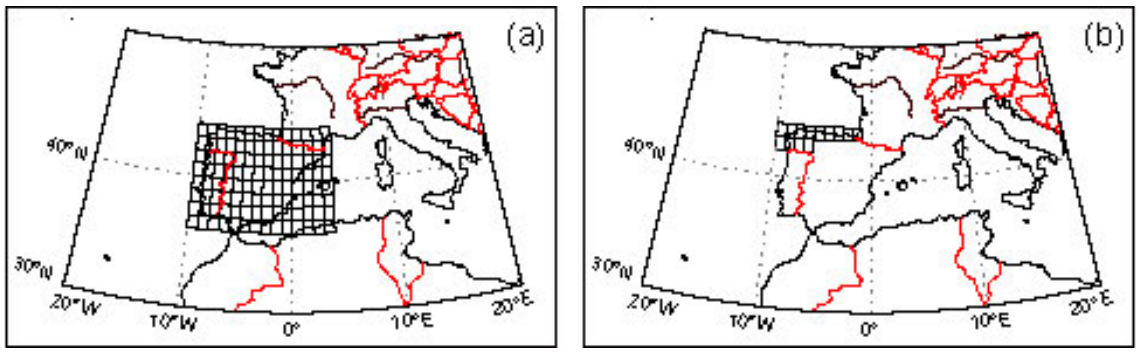


Figura 2: (a) Rejilla peninsular 1ºx1º; (b) Rejilla limitada al área de una cuenca particular (la Cuenca Norte).

2.2 Registros Históricos de Observatorios

En lo que respecta a los datos de observaciones locales, hemos considerado la precipitación (Pp), la racha máxima de viento (Rx), y el % de insolación (In) en una red de l12 estaciones completas del INM que presentan un número de datos apropiado para el análisis (ver figura 3).

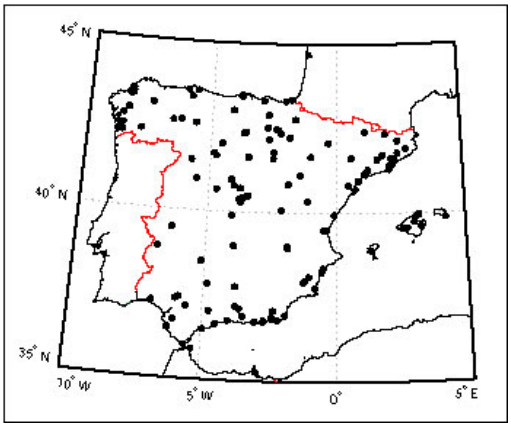
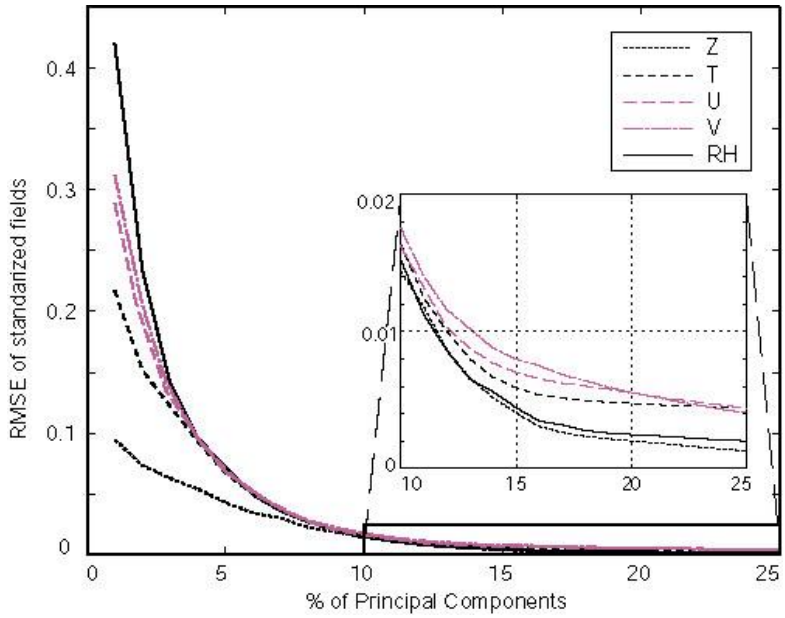


Figura 3: Red de estaciones completas del INM considerada para el estudio.

2.3 Eliminación de redundancia. Análisis de Componentes Principales.

El análisis de Componentes Principales (CPs) es una herramienta estadística eficiente para representar el máximo de varianza de un conjunto de datos con la menor dimensión posible (consultar, por ejemplo, Preisendorfer y Mobley, 1988). La dimensión de los vectores de estado de los modelos anteriores mostrados en [1] y [2] respectivamente, es muy elevada y contiene una gran redundancia espacial y temporal. Por tanto, antes de realizar cualquier cálculo resulta conveniente reducir esta dimensión minimizando la pérdida de precisión en la especificación de los campos atmosféricos; para ello nos hemos fijado un error máximo del 1% y hemos aplicado el método de componentes principales para obtener la reducción de dimensión correspondiente. La figura 4 muestra el Root Mean Square Error (RMSE) entre los campos 3D reales de cada una de las variables y los campos reconstruidos con el número indicado de CPs (utilizando una cierta proporción del número total de variables). Los cálculos mostrados corresponden al modelo M1, pero análogos resultados han sido obtenidos para el modelo M2. A partir de esta figura puede observarse cómo un 10% de las dimensiones es capaz de reproducir los campos con el error requerido. Por ello, en nuestro caso tomamos 500 CPs de los vectores de estado para llevar a cabo la



búsqueda de análogos.

Figura 4: RMSE de la reconstrucción de los campos para las variables indicadas considerando diversas dimensión para las CPs (hasta un 25% de la dimensión original del sistema).

3. Métodos de Clasificación para la Búsqueda de Análogos

El método estándar para la búsqueda de un conjunto de vectores análogos a un vector dado es el método de *k-vecinos*, que consiste en seleccionar el conjunto de los *k* vectores más próximos al vector dado según una métrica prefijada (normalmente la métrica Euclídea); en este método es necesario especificar el número de elementos *k* que compongan los conjuntos de análogos. Un método más sofisticado que realiza una agrupación en clases del espacio de estados del reanálisis es el método conocido como *k-medias*. En este caso el parámetro *k* indica el número de clases que se desea hallar; cada clase será el conjunto de análogos de los elementos contenidos. Una ventaja de este método es que produce conjuntos de análogos de tamaños variables, dependiendo de las concentraciones de patrones en el espacio. Junto con estos dos métodos también consideraremos una variación de éste último conocida como *redes auto-organizativas* (SOM, ver Kohonen 1995), que incluye una noción de vecindad entre clases o grupos próximos añadiendo una componente para preservar la topología en el algoritmo de clasificación. Por cuestiones de espacio obviamos la descripción de estos métodos (ver Cano 2001).

4. Validación de los Métodos

Con el fin de comparar los métodos anteriores realizamos una serie de experimentos de validación considerando la pericia de los modelos M1 y M2 separadamente en cada una de las cuencas peninsulares. Para ello se definieron distintos umbrales para la precipitación (Pp>0.5mm, 2mm, 10mm y 20mm), la racha máxima de viento (Rx>50km/h, 80Km/h) e insolación (In>20%, 80%). Las diferencias obtenidas entre los modelos no fueron demasiado significativas en promedio; sin embargo el método k-vecinos fue el que presentó una mayor varianza en los errores en casi todos los casos. En cuanto a los métodos de k-medias y de redes auto-organizativas, no se encontró ninguna diferencia general entre ambos. Por ello, dado que el método de k-medias es más eficiente desde un punto de vista computacional, resulta el método de análogos más conveniente de los analizados en este estudio (queremos hacer notar que el método de redes auto-organizativas posee otras ventajas que no se describen en este artículo, sobre todo de cara al tratamiento de predicciones por conjuntos EPS, ver Cano 2001).

Los resultados que se muestran a continuación corresponden al método de k-medias. En las siguientes tablas se indica el Brier Skill Score (BSS) obtenido en las distintas cuencas para D+1 utilizando los modelos M1 y M2 para las cuatro estaciones del año 1999 (Invierno: DEF, Primavera: MAM, Verano: JJA, Otoño: SON) (ver Brier, 1950 para más detalles sobre la validación de predicciones probabilísticas). El BSS de cada una de las estaciones, *j*, de una cuenca dada se define como:

$$BSS = 1 - \frac{BS_P}{BS_C}, \text{ donde } BS_P = \sum_{i=1}^{90} (P_{ij} - \hat{P}_{ij})^2 \ ;$$

y BS_C corresponde al Brier Score de la climatología estacional; P_{ij} y $\hat{P}_{ij}$ son respectivamente las probabilidades ocurridas y predichas de que el meteoro dado supere un cierto umbral el día *i* en la estación *j*. En las tablas que se muestran a continuación los valores indicados son BSS promedio de todas las estaciones de la cuenca indicada. Para mayor detalle sobre las validaciones se puede consultar la página `http://etsiso2.macc.unican.es/~meteo`.

	Pp								Rx				In			
	>0.5mm		>2mm		>10mm		>20mm		>50km/h		>80km/h		>20%		>80%	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
DJF	0,45	0,57	0,44	0,55	0,28	0,40	0,18	0,29	0,33	0,37	0,33	0,44	0,46	0,47	0,24	0,36
MAM	0,45	0,54	0,41	0,48	0,27	0,34	0,16	0,16	0,32	0,36	0,32	0,28	0,20	0,22	0,33	0,29
JJA	0,33	0,38	0,32	0,36	0,20	0,24	0,18	0,29	0,19	0,23	0,85	0,67	0,15	0,16	0,28	0,29
SON	0,42	0,54	0,35	0,48	0,29	0,36	0,23	0,25	0,33	0,36	0,39	0,47	0,28	0,27	0,24	0,25

Tabla 1: Resultados de la validación (D+1) en la *cuenca Norte* para los umbrales de precipitación, racha máxima de viento, e insolación indicados en el texto.

	Pp								Rx				In			
	>0.5mm		>2mm		>10mm		>20mm		>50km/h		>80km/h		>20%		>80%	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
DJF	0,64	0,70	0,56	0,60	0,32	0,33	0,31	0,29	0,37	0,37	0,37	0,42	0,34	0,40	0,43	0,46

MAM	0,31	0,41	0,14	0,27	0,12	0,28	-	-	0,19	0,26	-	-	0,21	0,23	0,34	0,42
JJA	0,18	0,23	0,42	0,72	0,57	0,90	0,32	0,91	0,14	0,15	0,91	0,80	0,45	0,66	0,32	0,40
SON	0,29	0,47	0,38	0,48	0,36	0,44	0,19	0,3	0,35	0,35	0,56	0,54	0,18	0,23	0,35	0,37

Tabla 2: Resultados de la validación (D+1) en la *cuenca del Guadalquivir* para los umbrales de precipitación, racha máxima de viento, e insolación indicados en el texto.

En las tablas anteriores se puede observar claramente que los mejores resultados se obtienen con el modelo M2, que limita el área a la región de interés, y cubre el rango horario asociado al meteoro que se desea predecir. Las diferencias son muy significativas en algunos casos con mejoras de hasta casi un 100% en la pericia.

Agradecimientos

Los autores agradecemos a la Universidad de Cantabria, al Instituto Nacional de Meteorología y a la Comisión Interministerial de Ciencia y Tecnnología (CICYT Proyecto REN2000-1572) su apoyo en este proyecto. Más información en: http://etsiso2.macc.unican.es/~meteo/

Referencias

Brier G.W. (1950): Verification of Forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

Cano, R., Cofiño, A.S., Gutiérrez, J.M. and Rodríguez, M.A. (2001): Self-Organizing Maps for Statistical Downscaling in Short-Term Forecasting. A Case Study of the Iberian Peninsula. Pendiente de publicación.

Gutiérrez J.M., Cano R., Rodríguez M.A., and Cofiño A.S. (1999) : Redes neuronales y patrones de analogías aplicados al downscaling en modelos climáticos, en Proceedings del II Congreso Nacional de Climatología, 234--241, Instituto Nacional de Meteorología, Madrid.

Kohonen, T. (1995): Self-Organizing maps. Number 30 in Springer Series in Information Sciences. \newblock Springer-Verlag.

Lorenz E. N. (1969): Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences, 26, 636-646.

Preisendorfer, R.W. y Mobley, C.D. (1988): Principal component analysis in meteorology and oceanography. Elsevier, Amsterdam.

Toth, Z. (1990): Estimation of atmospheric predictability by circulation analogs. Monthly Weather Review, Vol 119, Nº 1, 65-119, January 1991.

Zorita E. and Storch H.V. (1997): A survey of statistical downscaling techniques. GKKS Technical Report No. 97/E/20, Geesthacht, Germany.

Zorita E. and Storch H.V. (1999): The analog method as a simple statistical downscaling technique: comparison with more complicated methods. Journal of Climate, 12, 2474-2489.